基于关键词和关键句抽取的用户评论情感分析

喻 影[1] 陈 珂[1,2] 寿黎但[1,2] 陈 刚[1,2] 吴晓凡[3]
(浙江大学计算机科学与技术学院 杭州 310027)[1]
(浙江省大数据智能计算重点实验室(浙江大学) 杭州 310027)[2]
(网易(杭州)网络有限公司 杭州 310051)[3]

摘 要 情感分析的一项主要研究任务是根据文档内容对其情感极性(即正类和负类)进行判断。在判断文档的情感极性时,不同的词语和句子具有不同的情感贡献度,因此如何从整个文档中准确地提取与情感分类更相关的词语和句子,从而提升分类性能,成为了一个重要问题。在有监督实验中,基于依存句法关系分析句子的逻辑结构,提取出了与表达情感更相关的词语进行加权,提高了分类性能。在半监督实验中,使用基于中文评论的关键句抽取和分类器融合算法,对整篇文档中包含更多情感词和总结意味的关键句进行了抽取,充分考虑了句子的情感词属性、位置属性、标点符号属性和关键词属性,并且使用分类器融合算法,让置信度最高的子分类器决定分类效果。在大众点评网和头条新闻的数据集上将所提算法与已有的经典算法进行对比,发现所提方法的性能更高,从而证明了基于依存句法分析的关键词抽取和基于特征的中文关键句抽取算法的有效性。

关键词 情感分析,依存分析,关键句抽取,半监督学习,协同训练
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/jsjkx.191000531C

Sentiment Analysis of User Comments Based on Extraction of Key Words and Key Sentences

YU Ying[1] CHEN Ke[1,2] SHOU Li-dan[1,2] CHEN Gang[1,2] WU Xiao-fan[3]
(College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)[1]
(Key Laboratory of Big Data Intelligent Computing of Zhejiang Province (Zhejiang University), Hangzhou 310027, China)[2]
(Netease (Hangzhou) Network Co., Ltd, Hangzhou 310051, China)[3]

Abstract One of the main task of sentiment analysis is to determine the polarity of a review whether it is positive or negative according to the content of the document. When determining the emotional polarity of a document, different sentences and words have different emotional contribution on the classification result, so how to extract more related words and sentences becomes an important problem. In the experiment of the supervised classification, this paper used the dependency syntactic analysis to extract the words which are more related to the emotion and can improve the classification effect. In the semi-supervised classification experiment, the key sentence extraction and the combining-classifier method based on the Chinese comments have been used. For key sentence extraction, the proposed approach takes the following attributes into account: sentiment attribute, position attribute, special word attribute and punctuation attribute. This approach extracts key sentences containing more emotional words and summary meaning, then uses combining-classifier method to make the sub classifier with the highest confidence to determine the classification effect. The results show that the performance of the proposed method is better than the baseline methods, which proves the validity of keyword extraction based on the dependency parsing and key Chinese sentence extraction algorithms.

Keywords Sentiment analysis, Dependency parsing, Key sentence extraction, Semi-supervised learning, Co-training method

---

到稿日期:2018-07-11 返修日期:2018-09-15 本文受国家重点研发项目(2017YFB1201001),国家自然科学基金项目(61672455),浙江省自然科学基金(LY18F020005)资助。

喻 影(1993-),女,硕士,主要研究领域为数据挖掘、情感分析等;**陈 珂**(1977-),女,博士,副教授,硕士生导师,CCF 会员,主要研究领域为时空数据库、数据挖掘以及数据隐私保护等,E-mail:chenk@zju.edu.cn(通信作者);**寿黎但**(1974-),男,博士,教授,博士生导师,CCF 会员,主要研究领域为空间数据库、数据挖掘、数据可视化等;**陈 刚**(1973-),男,博士,教授,博士生导师,主要研究领域为大数据管理;**吴晓凡**(1984-),男,博士,主要研究领域为大数据智能。

1 引言

随着互联网的普及,越来越多的社交多媒体及服务网站开始出现,给广大用户提供了对产品以及服务进行反馈的平台,从而帮助产品及服务提供者更好地把握用户心理,提高产品质量。情感分析,能帮助研究者们从海量的文本中提取有效的用户观点和情感态度。根据评论内容对情感极性进行正负判断,是情感分析中一个比较重要的问题。本文主要从关键词和关键句抽取这两个角度对情感分类进行研究,致力于提高情感分类的性能。

本文的主要贡献如下:

(1)设计了一套基于关键词和关键句提取的情感分析框架,其包括:在基分类器的构建过程中,使用关键词提取和加权算法对句子中的情感因子进行增强;在半监督协同训练过程中,采用关键句抽取算法将句子集合划分为关键句集合和非关键句集合;在类别标注过程中,使用分类器融合的方法对上述过程得到的多个分类器进行集成。

(2)提出了一种基于依存句法分析的情感词抽取算法,主要通过分析句子的逻辑结构,结合词性标注,抽取与情感色彩更相关的词语进行加权。

(3)提出了一种基于中文评论的关键句抽取和分类器融合算法。在半监督过程中考虑句子的情感、关键词、位置和标点符号属性,将原始文档划分为不同且独立的两部分用于协同训练,并且在最后的类别标注中使用多个分类器融合的方法进行标注。

(4)在大众点评和头条新闻的数据集上,与已有方法进行对比,通过结果发现本文提出的算法框架的分类效果有了明显的提升。

2 相关工作

文本情感分析主要是通过使用自然语言处理和文本挖掘等技术来识别和提取文章中所表达的主观信息。一般而言,情感分析的一项重要任务就是判断文本中表达的情绪是积极的还是消极的。按照学习方法,情感分类可以分为无监督学习、有监督学习和半监督学习。

在无监督的学习方法中,所有数据都缺少标注,相关研究相对较少。其中,Turney[1]提出了一种基于文本词性标注的情感分类方法。这种方法分别计算句子中的候选词语与积极情感词、消极情感词之间的交叉熵,选取最大熵值对应的情感类别作为情感分类结果。

在有监督的情感分类问题中,主要从特征选择和分类器比较这两个角度进行分类效果的提升。针对特征选择问题,文献[2]首次引入了句子的语法结构(如依存关系)作为特征;文献[3]通过对文档进行自分割来提取关键信息;文献[4]对 n 元特征的不相关性和冗余性进行了过滤;文献[5]在 Twitter 数据上引入了词性特征和树核特征来进行表示;文献[6]则从语义角度出发,基于 WordNet 寻找中文词汇的相似词;文献[7-8]通过深度学习挖掘词之间的潜在语义;文献[9]采用跨领域的文本表情符号特征来协助文本分类;文献[10-12]使用更复杂的网络结构挖掘词语之间更深层次的特征;文献[13]结合外部情感词典的方法提取评论中的情感词特征;文献[14]结合语言本身的符号特征和词嵌入特征来挖掘更丰富的语义信息。上述这些特征选择和加强算法都在一定程度上改善了情感分类的效果。此外,常用的文本分类的有监督学习方法,包括 SVM、LR、Bayes、DT、KNN、Bayes、线性分类、决策树及 k-NN 等方法,模型简单,运行效率高,同时可以结合数据分布来有效提升文本分类的性能。集成学习也通过决策优化等手段将各个弱分类器进行综合,优化了分类系统的总体性能。决策优化的手段包括分类器投票、累乘取最大等,最终确定整个系统的输出类别。

半监督学习(Semi-Supervised Learning,SSL)使用大量的未标记数据以及少量的标记数据来进行模式识别工作。目前,互联网数据呈现爆发式增长,数据的标注需要大量的人工工作,同时准确率也受到质疑,因此半监督学习正越来越受到人类的重视。在半监督分类学习中,目前使用最为广泛的学习方法有:自训练、基于差异的方法、生成式方法、判别式方法和基于图的方法等。自训练算法是 Scudde[15],Fralick[16] 和 Agrawala[17] 提出的最早将未标注样本应用于半监督训练中的算法。其首先使用少量未标注的样本来训练初始分类器,然后使用该分类器对未标注数据进行标注,再按照分类的置信度进行排序,选出分类准确度最高的几个样本并对其进行标注。当达到迭代轮数或者所有未标注样本都已经完成标注后,停止迭代。与自训练算法对应的是协同训练算法[18-19],其被看作"多视图"学习,利用了多视图的"相容互补性",假设数据拥有两个充分且条件独立的视图。"充分"是指每个视图都包含足以产生最优学习器的信息;"条件独立"是指在给定类别标记的条件下两个视图独立,并且利用未标记数据进行训练。在每个视图上,先基于已标记的数据训练出一个分类器,然后让每个分类器去选择置信度最高的样本赋予伪标记,最后将伪标记样本提供给另一个分类器作为新增的有标记样本用于训练更新。这个"互相学习,共同进步"的过程不断迭代进行,直到两个分类器都不再发生变化。协同训练算法本身是为多视图数据设计的,但是在单视图数据上也可以实现,其中视图之间的差异包括不同的学习算法、数据采样或者参数设置。

3 研究框架

3.1 基于关键词和关键句提取的情感分类框架

本文主要考虑从关键词和关键句这两个角度入手,通过提取与情感表达更相关的词语和句子,来提升情感分类的效果。具体的算法流图如图 1 所示,具体可以分为关键词加权算法、关键句抽取算法和分类器融合算法这 3 个部分。

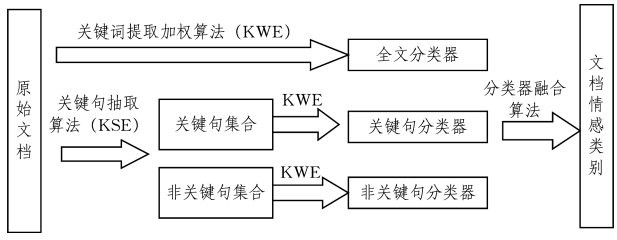

图 1 基于关键词和关键句提取的算法流图

Fig. 1 Flowchart of the algorithm based on extraction of key words and key sentences

首先,在单独的基分类器设计中使用依存句法对句子的逻辑结构进行分析,并且进行词性标注,从而在指定的依存关系对中抽取出与情感表达更相关的形容词、副词和动词,比如表达情感程度深浅的"非常"、表达对事物推崇行为的"推荐"和对事物欣赏态度的"青春"等,再通过具体实验确定这些词语的加权系数,结合特征选择算法进行词语过滤和选择。后续的相关实验表明,对关键的情感特征进行加权会引入更多相关的有用信息,从而使得基分类器的分类效果得到提升。

其次,由于本次实验的数据集中包含大量的未标记数据以及少量的标记数据,因此考虑使用半监督学习中的协同算法进行训练。在协同训练算法的视图划分中,本文设计了一种关键句抽取算法,主要考虑了句子的 4 个特征,包括情感词属性、位置属性、标点符号属性和关键词属性,将原始文档划分为不同且独立的两部分,从而训练出关键句分类器、非关键句分类器和全文分类器这 3 个分类器。

最后,设计了一种简单的分类器融合算法,让置信度最高的子分类器决定分类效果,而不是采用单一分类器去对测试文档的标签进行标注。

后续的相关实验表明,不管是关键句抽取算法还是基于最大置信度的分类器融合,都可以提高情感分析的分类效果。

3.2　基于依存句法分析的情感词加权算法

本节提出了一种基于依存句法分析的情感词加权算法,其使用依存句法分析句子的逻辑结构,并从中抽取出与情感表达更相关的词语,对这些词语进行加权,修正特征权重,最后建立分类模型。

使用依存句法分析不仅可以对句子中的词语进行切分,还可以进行词性标注以及依存关系对的提取。一般而言,副词、动词和形容词常被用来表达观点和认知,如"非常"可以用来表示情感程度的深浅,"推荐"可以用来表示对事物的赞赏和推崇行为,"漂亮"可以用来表示对事物的一种欣赏的态度。因此,可以结合依存关系和词性标注,根据表 1 所列规则,从评论中提取出与情感表达更相关的词语。

表 1　利用依存关系提取情感词对

Table 1　Extracting combination of emotional words using dependency relationship

规则	词性链	依存关系	示例
1	$adj+n$	att	性价比高,系统流畅
2	$adv+adj$	adv	很可爱
3	$adv+vt$	vtadvmod	很喜欢
4	$adv+n$	npadvmod	很青春
5	$vt+n$	vob	推荐这款

对提取出的依存关系对中的副词和形容词进行加权。经典的 TF-IDF 的计算公式如式(1)所示:

$$w_{i,j}=tf_{i,j}*idf_i \tag{1}$$

为了提高某些词语的权重,引入权重因子 K 对一些词语的频率进行加强。改进之后的权重计算公式如式(2)所示:

$$w_{i,j}=K*w_{i,j}=K*tf_{i,j}*idf_i \tag{2}$$

结合词性标注的结果,对形容词、副词和动词分别采用不同的权重进行加权,具体的取值将在第 4 节的实验部分进行介绍。

3.3　关键句抽取算法和分类器融合

本节设计了一种基于中文关键句抽取和分类器融合的情感分类方法。将一个评论文本中的相关句子划分为关键句和非关键句,其中关键句处于文本中的突出位置,含有更多具有情感色彩的词语,同时具有一定的总结性;非关键句则内容相对更长,含义更复杂。在得到句子的划分后,结合 co_training 协同训练的相关方法对这部分特征进行实验。

在关键句和非关键句上训练得到的分类器各有特色,且相互补充。关键句集合上的词汇分布相对集中,含有更多包含情感色彩的情感词,如"喜欢""推荐"等,如果使用基于关键句的分类器进行分类,会得到很好的分类效果;另一方面,非关键句分类器则会对关键句分类器进行补充,提供更多的细节信息,而且更多的词汇特征可能会带来更好的分类效果。

关键句抽取算法主要考虑了 4 个属性,分别是情感词属性、位置属性、关键词属性和标点符号属性。一个理论意义上的关键句应该包含更丰富的情感词,这些情感词可能位于评论文本的首尾位置以强调感情,包含一些总结性的词语,同时拥有一些增强感情色彩的标点符号。

给定一条评论,对其中的每个句子分别计算 4 个属性的得分,然后进行加权求和,其中 $f(s_i)$ 得分最高的句子则被确定为情感关键句。具体的形式化定义如下:若任意评论 d 由一系列的句子组成,即 $d=\{s_1,s_2,\cdots,s_m\}$,其中 m 代表句子数目,而每个句子 s_i 则由一系列词语组成,每个句子的最终得分可以表示成式(3)中的 4 个属性的加权求和 $f(s_i)$。

$$f(s_i)=\lambda_1*f_{emotion(s_i)}+\lambda_2*f_{position(s_i)}+\lambda_3*f_{keyword(s_i)}+\lambda_4*f_{punctuation(s_i)} \tag{3}$$

其中,$\lambda_1,\lambda_2,\lambda_3,\lambda_4$ 分别是情感词属性、位置属性、关键词属性和标点符号属性的权值,具体权值的大小由最大化分类器的精度确定。

(1)情感属性

关键句一般通过使用更多的情感词(动词、副词、形容词等)来表达作者的观点,因此在评价一个句子是否具有情感属性并衡量情感值的强弱时,使用如下计算方式。本文采用的情感词典由《知网情感词典 HowNet》和《台湾大学简体中文情感极性词典 NTSUSD》整理得到。

根据词性标注的结果,将句子中的情感词分为两部分,分别为动词 vt 和〈副词,形容词〉组合对〈adv,adj〉。对于动词部分的感情色彩,按照情感得分进行累加;而〈副词,形容词〉部分则使用依存句法提取出的 ADV 组合对,对副词的程度得分和情感词进行乘法运算从而得到情感得分,最后将两部分的情感得分值进行累加。

$$f_{emotion(s_i)}=\frac{\left|\sum_{type(w_{ij})=vt}w_{ij}+\sum_{pair_{ij}=\langle adv,adj\rangle}pair_{ij}\right|}{n} \tag{4}$$

其中,w_{ij} 为使用词性标注后句子中的动词部分的情感得分,adv_{ij} 为副词的情感程度得分,adj_{ij} 为形容词的情感得分。

$$w_{ij}=\begin{cases}+1, & w_{ij}\in pso\_dict \text{ and } type(w_{ij})=vt\\ -1, & w_{ij}\in neg\_dict \text{ and } type(w_{ij})=vt\\ 0, & w_{ij}\notin pso\_dict\cup neg\_dict\end{cases} \tag{5}$$

$$pair_{ij}=adv_{ij}*adj_{ij} \tag{6}$$

(2)位置属性

对于评论而言,根据写作的一般特点,大多数作者会习惯于在文本首尾发表和强调自己的观点。因此,一般而言,位于新闻评论头部和尾部的句子成为关键句的可能性更大。

$$f_{position(s_i)}=\begin{cases}1, & \text{if } i==0 \text{ or } i==len(s)-1\\ 0, & \text{if } 0<i<len(s)-1\end{cases}$$

(3)关键词属性

情感分析的关键句,通常会包含一些总结性和针对性的词汇或者短语,用于对整个文本进行总结。因此,基于整个新闻评论样本,对每篇评论的最后一句话中各个词语出现的次数进行统计,并根据词频进行排序,从而得到中文评论中常见的总结性关键词,如表 2 所列。

表 2 中文评论中常用的 Top10 词汇

Table 2 Top10 frequent words in Chinese reviews

序号	词语	序号	词语
1	总体	6	但是
2	所以	7	一直
3	因此	8	最后
4	总之	9	反而
5	总的来说	10	基本上

倘若这些关键词出现在某一句子中,则将该句定义为关键句,具体的规则如式(7)所示:

$$f_{keywords(s_i)}=\begin{cases}1, & \text{if } word_{ij}\in keywords \text{ and } word_{ij}\in s_i\\ 0, & \text{if } word_{ij}\notin keywords \text{ and } word_{ij}\in s_i\end{cases} \tag{7}$$

(4)标点符号属性

在评论中,发表者一般会使用问号、感叹号、波浪号等来对情感进行增强和渲染,以表达自己的反问、愤怒、开心或者满意的心情。因此,倘若这些句子出现在某一句子中,则将该句定义为关键句,具体的规则如式(8)所示:

$$f_{punctuation(s_i)}=\begin{cases}1, & \text{if } word_{ij}\in\{?,!,\sim\} \text{ and } word_{ij}\in s_i\\ 0, & \text{if } word_{ij}\notin\{?,!,\sim\} \text{ and } word_{ij}\in s_i\end{cases} \tag{8}$$

在使用关键句抽取算法后,会将同一条新闻评论样本划分为关键句集合和非关键句集合,它们从不同的角度提供了关于评论情感分析的相关信息。一般而言,关键句处于评论的关键位置,情感信息丰富,能够提供更多、更关键的分类信息;非关键句集合则可能包括一些对情感分类效果不明显的噪音特征,或者从细节角度提供一些补充。

本节一共训练了 3 个 LR 基分类器,其中用关键句集合训练出分类器 f_1,用非关键句集合训练出分类器 f_2,同时用所有的评论句子训练出全文分类器 f_3。

在使用分类器对测试集中的样本进行预测时,基于情感因子加强的 LR 基分类器不仅可以为每个评论样本指定类别,还可以给出其所属类别的概率大小。因此,我们的融合策略从如下角度为评论样本进行了类别判断。假设拥有最高分类置信度的分类器包含最有效的用于情感分类的关键特征,这一分类器可能是关键句分类器 f_1,也可能是包含更多细节或冗余信息的非关键句分类器 f_2,还可能是包含最全面信息的全文分类器 f_3。这个假设是合理的,因为有效特征(表达情感的特征)通常与类别之间的联系更紧密,从而会有更高的分类置信度。因此,本文的融合算法中每个未标注样本的类别由置信度最高的子分类器决定,通过这种方法来避免噪音特征在分类决策中的不良影响。整个融合策略得到的最终类别标签 j 可以用式(9)表示:

$$\begin{aligned}&p(c_0\mid d)=\max_{i=1,2,3}(f_i(c_0\mid d))\\&p(c_1\mid d)=\max_{i=1,2,3}(f_i(c_1\mid d))\\&j=\begin{cases}0, & \text{if } p(c_0\mid d)>p(c_1\mid d)\\ 1, & \text{if } p(c_1\mid d)>p(c_0\mid d)\end{cases}\end{aligned} \tag{9}$$

其中,c_0 代表情感分类结果的极性为负,c_1 代表情感分类结果的极性为正。

Co_training 算法最早由 Blum 和 Mitchell 提出,是一种 Bootstrap 模式的学习方法,需要数据的两组特征集能够充分表示数据并且相互独立,一般会在两组不同的特征集下分别训练分类器。上面提出的关键句抽取算法刚好将评论文本划分为独立又不影响的两部分,而分类器融合则通过将不同的分类器进行融合,来综合得到分类结果。因此,采用半监督学习中的 Co_training 算法来进行新闻评论的情感识别,分别训练关键句分类器、非关键句分类器和全文分类器。为了保证在算法过程中加入文本的准确性,采用协商策略[4],在每一个步骤中在选择最可能加入的样本时,不是把多个角度学习获得的所有正例和负例进行取并操作,而是选择从各个角度来看都是最有可能的结果,即对分类器结果进行取交操作。具体的算法流程如算法 1 所示。

算法 1 基于中文关键句抽取的 Co_training 算法

Input:Feature set=$\{fea_{key},fea_{not\_key},fea_{all}\}$,training set L,testing set U

Output:Sentiment labels of testing set

1. while iteration<N:
2. f_1<-selected features from fea_{key} and to train a classifier
3. $probs_i$<-predict U by f_1
4. S_1<-sorted{$probs_1,\cdots,probs_i$} and selected top K items
5. f_2<-selected features from fea_{not_key} and to train a classifier
6. $probs_j$<-predict U by f_2
7. S_2<-sorted{$probs_1,\cdots,probs_j$} and selected top K items
8. f_3<-selected features from fea_{all} and to train a classifier
9. $probs_k$<-predict U by f_3
10. S_3<-sorted{$probs1_1,\cdots,probs1_k$} and selected top K items
11. U′<-U-($S_1\cap S_2\cap S_3$)
12. L′<-L∪($S_1\cap S_2\cap S_3$)
13. f<-selected f from $\{f_1,f_2,f_3\}$ which has the highest scores
14. predict test by f

4 实验与分析

4.1 实验数据及流程

为了考查算法的鲁棒性,本文设计爬取了两套数据集,一套来源于大众点评餐馆的用户评论数据集,另一套来源于头条新闻各品牌的用户评论数据集。

具体的数据规模及条目如下:

(1)大众点评网上爬取的餐饮类别包括川菜、海鲜、湖北菜、韩式料理、面包甜点等,一共 200 013 条用户评论,涉及

193 家餐馆。采用众包的多数表决法,人工标记了 8000 条样本,情感极性的正负比例为 1∶1。下面分别给出某餐厅在线评论中的正面情感评论和负面情感评论。

1)正面评论:金华的海底捞开了这么久,终于得空去吃了一次,那天去得早一下子就坐上了,服务员很热情,一来就送上毛巾水果手套啥的,点餐也是一台 ipad,海底捞的爆米花、妙脆角和番茄锅还是强烈推荐啊!

2)负面评论:价格太贵了,等了半天才等到位置,感觉菜品也不够新鲜,什么黄喉之类的,吃在嘴里酸酸的～～～～～希望大家不要被盛名骗了。

(2)头条新闻上爬取的品牌类目涉及美妆美容、电子产品、运动服饰等多个类别,一共 137680 条用户评论,涉及 401 家品牌。采用众包的多数表决法,人工标记了 8000 条样本,情感极性的正负比例为 1∶1。下面分别给出某手机评论中的正面情感评论和负面情感评论。

1)正面评论:支持华为买中国手机!

2)负面评论:买了款 mate8 电池用了 13 个月就疲软了,找售后说退厂换电池得 40 天,还牛逼哄哄的,用不起不用了。

整个实验主要分为两部分:1)验证半监督学习中构建基分类器时使用的关键词抽取算法的有效性,具体包括关键词加权前后的分类效果对比和加权系数的确认;2)验证半监督学习中进行视图划分时使用的关键句抽取算法和分类器融合算法的有效性,具体包括确定每次加入的样本数目和基于不同的标注训练集大小研究精确度、召回率和 F1 这 3 个分类效果指标的变化。

4.2　关键词抽取实验

3.2 节提出使用依存句法分析的方法来对评论中的关键情感词的特征权重进行增强,并且构建了一个情感词加权的 LR 基分类器。使用标注好的文本数据,通过相关对比实验,从分类准确率、召回率、宏平均值和几何平均值等方面来验证该方法的有效性。

实验 1　未对情感词进行加权的特征保留百分比

使用卡方检验的方法进行特征选择时,将每个词语的 TF-IDF 值作为特征权重,并对文本进行向量化。在向量化之后,使用 chi2 的特征选择方法对数据进行降维,以解决维数灾难问题,去除不相关特征,减少模型训练时间并提高预测准确度,其中特征的保留比例会影响最后的实验结果。使用 5 折交叉检验的方法,设置不同的特征保留比例,利用几何平均值 *G_Mean* 和平均 *F*1 值来对模型的效果进行评估。基于大众点评网的具体实验结果如图 2 所示。

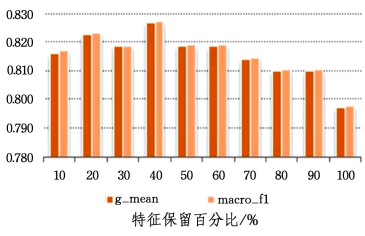

图 2　未加权的不同特征保留百分比实验(大众点评数据集)

Fig. 2　Unweighted character experiments based on different retentions (Dazhongdianping Dataset)

通过图 2 发现,当特征保留比为 40%时,情感分类的效果最佳,平均 *F*1 值最大,为 0.827;另外,随着保留特征数目的增加,分类的效果先提升后下降,尤其是保留全部特征时,相比只保留 10%的特征,分类效果相差 2%。以上数据说明:当特征数目增加时,不一定会带来分类效果的提升,可能还会引入很多无用的特征;同时,在采用 TF-IDF 值作为特征权重的情况下,保留较少的特征,反而会取得较好的预测效果。

基于头条新闻的数据集的实验结果同样显示,更多的特征保留比不一定会带来分类效果的提升。其中,特征保留比为 30%时,分类效果最佳,*F*1 值为 0.786;而当特征保留比为 90%时,*F*1 值降到了 0.747,大约相差 3 个百分点。

实验 2　引入情感词加权的特征保留百分比

在使用依存句法分析的方法进行句子逻辑结构的整理后,抽取出一个个重点关系对,并结合情感词典方法对与表达情感更相关的副词、形容词和动词做加权后,选择不同的特征保留百分比。基于大众点评餐馆数据集进行实验,结果如图 3 所示。

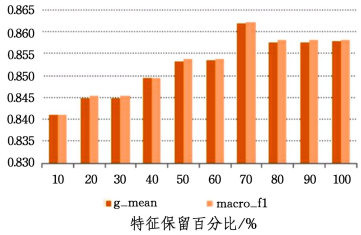

图 3　已加权的不同特征保留百分比实验(大众点评数据集)

Fig. 3　Weighted character experiments based on different retentions (Dazhongdianping Dataset)

当特征保留比为 70%时,情感分类的效果最佳,宏平均 *Macro_F*1 值最大,为 0.862。另外,保留特征数目的增加,普遍带来了分类效果的提升。相比只保留 10%的特征,保留 70%的特征将分类效果大约提升了 2 个百分比,说明在采用情感词加权算法后,对与情感表达更相关的词语进行了加权,从而在进行特征选择时,越大的特征保留比通常会引入更多提升分类效果的词语,方法的有效性得到验证。

通过表 3 发现,在利用依存句法分析并且结合词性标注进行情感关键词的抽取和加权后,分类效果得到了提升。其中,最佳特征保留百分比为 70,平均 *F*1 值为 0.863,几何准确率 *G_mean* 值为 0.862,相比未加权的实验 1,分类效果提升了 3.6%。

表 3　是否加权的情感分析实验结果的对比(大众点评数据集)

Table 3　Comparison of experimental results based on weighted sentiment analysis (Dazhongdianping Dataset)

	最佳特征保留百分比	平均 *F*1 值	几何准确率 *G_mean* 值
实验 1:未加权	40	0.827	0.828
实验 2:加权	70	0.863	0.862

使用头条新闻数据集进行实验的结论与图 2 一致(见表 4),对有效情感特征进行加权后,越多的特征保留比一般会带来更多分类效果的提升。通过实验发现,当特征保留比为

80%时，$F1$ 值最高为 0.812；而当特征保留比为 10%时，$F1$ 值仅为 0.773。相比未加权的实验 1，分类效果提升了 2.6%。

表 4 是否加权的情感分析实验结果的对比(头条数据集)

Table 4 Comparison of experimental results based on weighted sentiment analysis (Toutiao Dataset)

	最佳特征保留百分比	平均 $F1$ 值	几何准确率 G_mean 值
实验 1：未加权	30	0.786	0.785
实验 2：加权	80	0.812	0.813

实验 1 和实验 2 的差别仅在于利用了相关技术，从语句的逻辑结构上抽取了与表达情感更相关的词语进行加权。从情感分类问题入手，对文本进行向量化之后，可以将特征分为两部分，一部分是与正负情感相关的特征，另一部分是相对无关的特征。相关特征越多，分类器越能从文本中得到更多有用的信息用于分类。但是在实验 1 中，直接使用 TF-IDF 值确定特征权重时，更多的特征保留比不一定代表更多的相关特征引入，反而可能会引入一些无关特征，从而影响分类效果；在实验 2 中，对关键的情感特征进行了加权，更多的特征意味着更多相关的有用信息，从而使得分类效果有了进一步的提升。

对比大众点评和头条新闻的实验结果可以发现，本文所提出的情感分析方法在不同的数据集上具有相似的表现，即对情感词加权之后会带来分类效果的提升，同时，一般而言，加权后特征保留比越大时分类效果会越好。头条数据集上的实验结果直观上并没有大众点评数据集上的实验结果好，从数据源角度看，这主要是因为头条数据集的用户评论相对较短，而且在表达情感上使用的词语比较分散。此外，对于关键词加权算法，使用依存句法分析后，根据规则进行关键词抽取，而新闻的评论一般更加口语化，规则性较弱，因此提取效果会受到一定影响。上述两点导致了头条数据集上的实验性能相对较弱。

实验 3 确定权重大小

选取特征保留比例为 70%，分别改变提取出的依存对中形容词、副词和动词的权重大小，记录分类结果的 F1 值，具体的实验结果如表 5 所列。其中，K_1 代表针对形容词和副词的权重，K_2 代表针对动词的权重，由于动词表现的态度倾向和行动倾向更明显，因此设置 K_1 比 K_2 小，使得实验结果呈现一个上三角的结果。从表 5 中发现，当 $K_2=1.3$，$K_1=1.2$ 时，模型分类的效果最佳，为 86.3%，相比不加权的分类算法，效果提升了 3.6%。

表 5 不同权重下的实验结果(大众点评数据集)

Table 5 Experiment results based on different weights (Dazhongdianping Dataset)

$K_1\backslash K_2$	1	1.1	1.2	1.3	1.4	1.5
1	0.835	0.831	0.833	0.837	0.832	0.836
1.1	—	—	0.847	0.839	0.842	0.851
1.2	—	—	—	0.863	0.858	0.861
1.3	—	—	—	—	0.860	0.842
1.4	—	—	—	—	—	0.846
1.5	—	—	—	—	—	—

4.3 关键句抽取实验

由于大多数评论文本缺少标注，考虑采用半监督的学习方法，进而提出了一种基于中文文本的关键句抽取算法。将每条评论文本中的相关句子划分为独立不重合的两部分，关键句集合提供更多情感色彩和总结强调的功能，非关键句集合则提供更全面或冗余的相关视角。本节基于实验的语料数据，与其余 baseline 方法进行对比，证明了该方法的有效性。

为了对本文所提算法的有效性进行检验，结合半监督学习中的经典算法，将本文算法与下面几种 baseline 算法进行对比。

(1)算法 1 为经典的 self-training 算法，以 2.1 节中的 LR 分类器为基础，每轮迭代选取最确信的 n 个样本加入训练集，在下文被简称为 self-FS。

(2)算法 2 为 Su 等[20]提出的基于特征子空间与协同训练的情感分类算法，每次迭代过程中将全部特征划分为两个独立不相交的子空间，然后分别训练两个分类器，每轮迭代中由这两个分类器来选择样本，最后利用集成的方法对多个分类器进行整合，对样本标签进行标注。这种基于 Random-View 的 Co-training 算法，在下文被简称为 Co-training-RV 算法。算法 2 的存在主要是为了检验分类效果的提升得益于正确关键句的抽取，即基于关键句的特征划分相比随机的特征划分会带来分类效果的提升。

(3)算法 3 为基于全文分类器的 Co-training 算法，依旧采用关键句抽取算法提取关键句，每轮迭代中由 3 个基础分类器来选择样本，但是与本文算法不同，由融合后的分类器来指定样本的最终标签。算法 3 的存在主要是为了检验分类效果的提升同样得益于分类器的融合，即综合考虑了 3 个分类器的融合算法会带来分类效果的提升。

在半监督学习中，标注数据集的大小会影响算法的实验结果。当标注数据全部被标注时，半监督学习会退化成监督学习；当标注数据的数量很少时，半监督学习会接近无监督学习。

一般而言，半监督学习过程中会存在错误累积现象，因此大部分的半监督学习会将初始的标注样本设置为一个较大的值，从而获得较好的实验结果。但是本文算法将关键句抽取和有策略的分类器融合，充分利用评论中最有效的优质特征进行分类，从而减小噪音特征的影响和错误累计，因此本次实验从标注训练集的大小出发，对算法的性能进行研究。

具体实验时，从标记好的 8000 条数据集中抽取 6000 条作为测试集，再从剩余的 2000 条数据集中依次抽取不同大小的数据作为初始标注样本。在每次迭代过程中选择 n 个样本，分别在 n 取 1，5，10，15，20，25 时进行实验，在几种情况下均能取得不错的效果，但在 n 取 10 时，效果最佳，因此设置 n 为 10。计算 4 种算法的 $precision$，$recall$，$F1$ 值的变化情况。

如图 4 所示，当标记的训练集大小为 80 时，本文提出的融合后的 Co-training 分类器的精确度要高于一般的 self-FS 和随机特征划分的 Co-training-RV 算法。基本上随着标注的训练集的增大，各种算法的分类精度都在增加。此外，当标注的训练集大小为 280 时，self-FS 和 Co-training-RV 算法的分类结果都出现了较大幅度的下滑，这主要是因为子特征空间划分算法 Co-training-RV 中每次特征划分都具有一定的随机性，而 self-FS 算法中种子数据的质量也不够高，因此鲁棒性明

显减弱;而基于关键句抽取的算法(不管是全文分类器还是融合之后的分类器算法),都在原基础上保持了一定程度的增长。

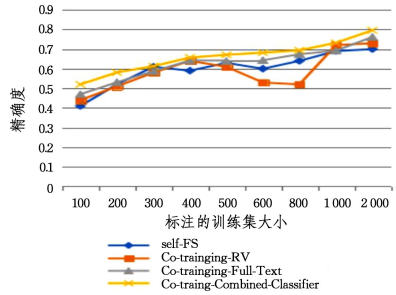

图 4　精确度随标注训练集大小的变化情况(大众点评数据集)

Fig. 4　Change of precision with size of tagged training set (Dazhongdianping Dataset)

从精确度指标来看,所提算法在不同大小的标注训练集上,基于自训练的 self-FS 算法有 0.6%～26%的提升,平均提升幅度为 10.9 %;基于随机特征划分的 Co-training-RV 算法有 0.9%～28%的提升,平均提升幅度为 13.7%。同样地,基于关键句抽取算法,融合后的分类器相比直接采用全文分类器,在精确度上大约会有 0.4%的提升幅度。

当标注的训练集增大时,4 种算法的召回率的变化情况如图 5 所示。相比于精确度而言,几种算法的召回率都有了一定程度的提升,并且相对平稳,但用于对比的 self-FS 和 Co-training-RV 算法的召回率仍然有一定波动。随着标注训练集的增大,不管是基于关键句抽取的 Full-Text 算法还是 Combined-Classifier 算法,都保持了平稳的增长,而其余两种算法在标注训练集大小为 200 和 320 时 recall 都出现了小幅度的下降。

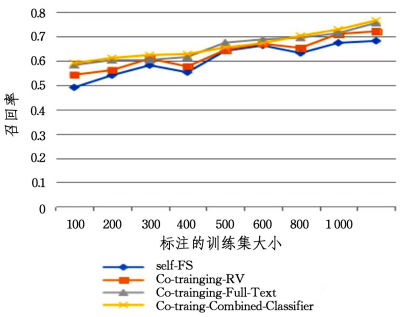

图 5　召回率随标注训练集大小的变化情况(大众点评数据集)

Fig. 5　Change of recall with size of tagged training set (Dazhongdianping Dataset)

从召回率指标来看,所提算法在不同的标注训练集大小上,基于自训练的 self-FS 算法有 1.5%～21.8%的提升,平均提升幅度为 9.98%;基于随机特征划分的 Co-training-RV 算法有 1.7%～8.9%的提升,平均提升幅度为 5.3%。同样地,基于关键句抽取算法,融合后的分类器与全文分类器在召回率上的表现比较类似,前者相比后者大约会有 0.13%的提升幅度。

图 6 给出了标注的训练集增大时,4 种算法的 F1 值的变化情况。F1 值与 precision,recall 相关,显示了算法的直接性能,可以用于不同算法的分类性能的对比。当标注的训练集大小为 80 时,本文提出的基于关键句抽取算法的 Combined-Classifier 的算法结果比 self-FS 高 10 个百分点。随着标注训练集的增大,不管是基于关键句抽取的 Full-Text 算法还是 Combined-Classifier 算法,都保持了平稳的增长,但是其余两种算法的 F1 在标注训练集大小为 200 和 320 时,都出现了一定程度的波动。

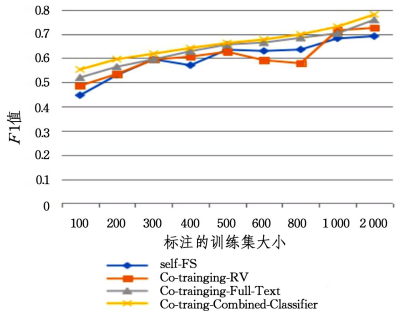

图 6　F1 值随标注训练集大小的变化情况(大众点评数据集)

Fig. 6　Change of F1 with size of tagged training set (Dazhongdianping Dataset)

综合上述 3 个指标,当标注的训练集较小时,本文提出的关键句抽取算法比其他两种基本算法的分类效果更好。其原因主要是,本文提出的基于关键句抽取和分类器融合的情感分类算法可以很好地识别出评论中的关键句,并且通过分类器融合的方法,充分利用与情感分类相关的优质特征进行分类,尤其是在标注训练集数量很小的情况下,噪声特征的引入不但对情感分类毫无帮助,而且会严重影响分类器的性能。因此,在训练样本较小的情况下,其他算法表现不佳,并且随着样本的扩大,会出现一定程度的波动;而本文的关键句抽取算法不仅在初始标注样本量较小时有较好的分类效果,而且随着标记样本的扩大,分类效果依旧会保持平稳上升的趋势。

结束语　本文引入了关键词和关键句这两个概念,通过有效地识别出与情感态度更相关的词语和句子,可以提高情感分类的性能;同时提出了一种基于依存句法分析的情感特征词抽取算法以及一种基于句内特征的关键句自动抽取算法。在有监督学习的基分类器设计中,本文通过分析句子的逻辑结构提取关键词,并且通过实验确定不同词性的关键词权重大小,从而使得在特征选择的过程中能引入更多与情感表达相关的词语特征,以提升分类的效果;在半监督的迭代过程中,采用关键句抽取算法,对关键句和非关键句进行划分,充分利用了关键句和非关键句两者之间的差异性和互补性,并且通过分类器融合算法选取置信度最高的分类结果进行标注,基于交集的 Co-training 算法使得两者可以互相学习、互相促进。最后,将所提算法与已有的经典方法进行对比,实验结果证明了本文方法的有效性。

参 考 文 献

[1]　TURNEY P D. Thumbs up or thumbs down?:semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002:417-424.

[2] NAKAGAWA T,INUI K,KUROHASHI S. Dependency tree-based sentiment classification using CRFs with hidden variables [C]// Human Language Technologies:Conference of the North American Chapter of the Association of Computational Linguistics. Los Angeles,California,USA,DBLP,2010:786-794.

[3] MCDONALD R T,HANNAN K,NEYLON T,et al. Structured Models for Fine-to-Coarse Sentiment Analysis[C]// Proceedings of the,Meeting of the Association for Computational Linguistics (ACL 2007). Prague,Czech Republic,DBLP,2007:30-32.

[4] ABBASI A,FRANCE S,ZHANG Z,et al. Selecting Attributes for Sentiment Classification Using Feature Relation Networks [J]. IEEE Transactions on Knowledge & Data Engineering,2011,23(3):447-462.

[5] AGARWAL A,XIE B,VOVSHA I,et al. Sentiment analysis of twitter data[C]// Proceedings of the Workshop on Language in Social Media (LSM 2011). 2011:30-38.

[6] LIU J W,LIU Y,LUO X Q. Semi-supervised learning method [J]. Journal of Computer Science,2015(8):1592-1617.

[7] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv:1301.3781.

[8] LE Q,MIKOLOV T. Distributed representations of sentences and documents [C] // International Conference on Machine Learning. 2014:1188-1196.

[9] ZHOU X,WAN X,XIAO J. Cross-lingual sentiment classification with bilingual document representation learning[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016:1403-1412.

[10] BOJANOWSKI P,GRAVE E,JOULIN A,et al. Enriching word vectors with subword information[J]. arXiv:1607.04606.

[11] PETERS M E,NEUMANN M,IYYER M,et al. Deep contextualized word representations[J]. arXiv:1802.05365.

[12] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv:1810.04805.

[13] TAN S,WANG Y,CHENG X. Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples[C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,2008:743-744.

[14] CAMBRIA E,PORIA S,HAZARIKA D,et al. SenticNet 5:Discovering conceptual primitives for sentiment analysis by means of context embeddings[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[15] SCUDDER H J. Probability of error of some adaptive pattern-recognition machines [J]. IEEE Transactions on Information Theory,1965,11(3):363-371.

[16] FRALICK S C. Learning to recognize patterns without a teacher [J]. IEEE Transactions on Information Theory,1967,13(1):57-64.

[17] AGRAWALA A K. Learning with a probabilistic teacher[J]. IEEE Transactions on Information Theory,1970,16(4):373-379.

[18] PARK S B,ZHANG B T. Co-trained support vector machines for large scale unstructured document classification using unlabeled data and syntactic information[J]. Information Processing & Management,2004,40(3):421-439.

[19] KIRITCHENKO S,MATWIN S. Email classification with co-training[C]// Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research. IBM Corp,2011:301-312.

[20] SU Y,JU S,WANG Z,et al. Semi-supervised sentiment classification with random feature subspace method [J]. Journal of Chinese Information Processing,2012,26(4):85-91.